**Information Retrieval**

**26 November 2019 (First MidTerm)**

**Name: Surname: Matricola:**

**Ex 1 [rank 4]** Given the three sets A = {1, 5, 8, 9}, B = {2, 3, 8}, C = {2, 5, 8} drawn from the Universe {0, 1, …, 10}. Show how you can use the Min-hash approach to evaluate the Jaccard similarity between every pair of those sets, based on a sketch of size 3 which is defined according to the permutations P1(x) = 4x+2 mod 11,

P2(x) = 2x mod 11, and P3(x) = 5x+1 mod 11.

**Ex 2 [ranks 5]** Giventhe following three adjacency lists of a Web graph, compress them according to the algorithm seen in class, by assuming that every list can copy from anyone before it:

4 🡪 7, 8, 13, 14, 18, 19

5 🡪 7, 9, 11, 13, 14, 18

6 🡪 7, 8, 11, 13, 14, 19

**Ex 3 [ranks 4+4]** Giventhe following files f\_old = “aaabb” and f\_new = “aabaaac”, show the execution of *rsync* and the execution of *zsync* on them by assuming a block of size 3.

**Ex 4 [ranks 5]** Given the dictionary of strings D = {aabb, aba, acac} construct a bigram index (hence k=2) and then search the string Q = “aabc” by assuming an edit-distance error e=1. Precisely, first use the overlap distance to filter a set of candidates, that are subsequently checked via dynamic programming, and return the string(s) at edit-distance 1.

**Ex 5 [points 4+4]** Answer to the following two questions:

1. Given two binary vectors of size d and having hamming distance h, state and prove the formula that defines the probability that LSH declares a match between them (the formula must use the parameters d and h).
2. Prove the error bound for the Spectral Bloom Filter.